Web spam szűrési módszerek

Doktori értekezés tézisek

Csalogány Károly

Témavezető: Benczúr A. András Ph.D.

Eötvös Loránd Tudományegyetem
Informatikai Kar
Információtudományi Tanszék

Informatika Doktori Iskola
Demetrovics János D.Sc.

Az informatika alapjai és módszerei doktori program
Demetrovics János D.Sc.

Budapest, 2009.

1. Bevezetés

A Web keresőrendszerei számára egyre nagyobb kihívást jelent a *Web spam*, a találati rangsor manipulációja céljából létrehozott tartalom terjedése [7]. Amit Singhal, a Google vezető kutatója becslése szerint 2004-ben a kereső spam iparágnak 4.5 milliárd dollár bevétele lehetett volna, ha képesek lettek volna befolyásolni az összes kereskedelmi jellegű keresések eredményeit [12]. A keresési listák előkelő helyei egyre nagyobb anyagi haszont jelentenek, így nem csoda, hogy a *spammerek* jelentős anyagi és emberi erőforrást fordítanak arra, hogy a keresőalgoritmusokat átverve befolyásolhassák a találatok sorrendjét. A jól ismert PageRank algoritmust többek között az is motiválta, hogy az egyszerű, csak a hiperhivatkozások számán alapuló rangsorolás nagyon könnyen befolyásolható volt.

A kereső spam tevékenység során nagy számú domain nevet kell regisztrálni és fenntartani, ami jelentős anyagi ráfordítást igényel. Az email spam működésmódjával ellentétben a spammerek egymással is versengenek a legjobb helyezésekért a keresési találatok között. A kereső spam tevékenység kőkemény üzlet az anyagi haszonért.

A spamet nehéz pontosan definiálni, sőt a becsületes és spam Weboldalak közti határt is nehéz meghúzni, így a spam szűrési módszereket hatékonyságát is nehéz mérni. Általában csak a találati listák minőségével való megelégedettséget tudjuk mérni, ami eléggé szubjektív.

Bár egyre bonyolultabb spam módszerek jennek meg, az alapvető technikák két kategóriába sorolhatók. A *tartalom spam* esetén a spammerek a Weboldal tartalmát módosítják, *link spam* esetén pedig az oldalak közötti hivatkozásokból alakítanak ki olyan struktúrát, ami félrevezeti a keresőalgoritmusokat.

2. Új eredmények

A Webes keresők üzemeltetőinek mindennapjaihoz hozzátartozik spam oldalak listáinak karbantartása, ami jelentős emberi erőforrásat igényel. Az új eredményeim célja ennek a munkának a megkönnyítése potenciális spam oldalak automatikus felismerésével, vagy ismert már spam oldalakhoz hasonló, valószínűleg spam oldalak felderítésével.

Értekezésemben öt új módszert írok le. Ebből kettő a Weboldalak körüli hiperhivatkozások elemzésére épül, a harmadik egyaránt használja link és tartalmi jellegzességeket, az utolsó kettő pedig a Weboldalak tartalma alapján ad módszert a Web spam szűrésére.

1. téziscsoport: SpamRank

A spammerek általában úgy próbálják növelni egy adott Weboldal PageRank értékét, hogy jelentős számú hiperhivatkozást hoznak létre, amelyek az adott oldalra mutatnak. Módszerünk automatikusan felismeri

Algoritmus 2.1 A SpamRank algoritmus vázlata.

for minden i Weboldalra **do**
  $\text{Support}_{i,\cdot} \leftarrow$ üres vektor
1. fázis: Támogatók keresése.
Olyan csúcsok generálása $\text{Support}_{i,\cdot}$-ba, amelyek jelentősen hozzájárulnak az i csúcs Pagerank-jéhez.
2. fázis: Büntetések kiszámítása.
for minden i Weboldalra **do**
  A Penalty_j büntetések kiszámítása $\text{Support}_{i,j}$-beli csúcsok PageRank-jének irregularitása alapján.
3. fázis: A SpamRank kiszámítása a büntetés vektoron értelmezett perszonalizált PageRank segítségével.
$\text{SpamRank} \leftarrow \text{PPR}(\text{Penalty})$

azokat a Weboldalakat, amelyeknek jogosulatlanul magas PageRank értékkel rendelkeznek. A módszer arra a feltételezésre épül, hogy a spameket támogató, jelentős forgalmat terelő oldalak PageRank értékének eloszlása eltérő a becsületes oldalakétól.

A SpamRank módszer fő elemeit a 2.1. Algoritmus foglalja össze. Az 1. fázisban (2.2. Algoritmus) minden Weboldalra kiszámolunk egy hozzávetőleges személyre szabott PageRank értéket, amelyhez a [J1]-ben leírt Monte Carlo közelítést használjuk. Az algorimus a Webgráf minden csúcsához körülbelül 1000 támogató csúcsot rendel a hozzájuk tartozó $\text{Support}_{i,j}$ értékekkel. Ez az érték annak a valószínűségnek felel meg, hogy a j csúcsból induló véletlen PageRank séta az i csúcsban ér véget. Másképpen azt fejezzük ki, hogy a támogató j csúcs milyen mértékben járul hozzá a i csúcs PageRank értékéhez.

Algoritmus 2.2 1. fázis: Támogatók keresése Monte Carlo szimulációval.

for minden j Weboldalra **do**
  for $\ell = 1, \ldots, N = 1000$ **do**
    $t \leftarrow$ véletlen érték egy ϵ paraméterű geometriai eloszlásból
    $i \leftarrow$ egy j-ből induló t hosszúságú véletlen séta végpontja
    $\text{Support}_{i,j} \leftarrow \text{Support}_{i,j} i + 1/N$

A SpamRank algoritmus 2. fázisában azonosítjuk azokat az oldalakat, amelyeket támogató csúcsok PageRank értéke eltér a megszokottól. A módszer az ismert Webgráf modellekre [1, 9] épül, amelyek szerint a Webgráfon a fokszámok hatványeloszlást követnek. A PageRank eloszlása nagyon hasonlóan viselkedik, és kísérletek [11, 5] azt mutatták, hogy nem csak az egesz Webgráfra érvenyes ez az eloszlás, hanem egy Weboldal kismeretű környezetében is.

Módszerünk arra a tapasztalatra épül, hogy a spam oldalak környezete jelentősen különbözik a becsületes oldalakétól, mivel sok, egymáshoz hasonló mesterségesen generált oldalt tartalmaz. Az eloszlás szabálytalanságának méréséhez megvizsgáljuk a hatványeloszláshoz illesztés hibáját a következőképpen. A támogató oldalakat vödrökbe osztjuk a Pagerank értékük alapján, úgy hogy az egyes vödrökhöz tartozó PageRank intervallum hossza exponenciálisan növekszik. Ha a PageRank értékek hatványeloszlást

Algoritmus 2.3 2. fázis: Büntetések kiszámítása.

Minden i-re $\text{Penalty}_i \leftarrow 0$
for minden i Weboldalra, amelynek legalább n_0, pozitív $\text{Support}_{i,j}$-vel rendelkező támogatója van **do**
  $\rho \leftarrow$ az i támogatóinak szabályossága
  if $\rho < \rho_0$ **then**
    for minden j Weboldalra, ahol $\text{Support}_{i,j} > 0$ **do**

$$\text{Penalty}_j \leftarrow \text{Penalty}_j + \begin{cases} (\rho_0 - \rho) & \{\text{1. változat}\} \\ (\rho_0 - \rho) \cdot \text{Support}_{i,j} & \{\text{2. változat}\} \end{cases}$$

    $\{\rho_0 = 0.85\}$
for minden j Weboldalra **do**
  if $\text{Penalty}_j > 1$ **then**
    $\text{Penalty}_j \leftarrow 1$

követnek, akkor az egyes vödrökbe kerülő oldalak számának a logaritmusa lineárisan növekszik. Az algoritmusunkban vödör sorszáma és az elemszám logaritmusa közötti Pearson korrelációt ($\rho \leq 1$) használtuk a szabályosság mértékeként, ahol a tökéletes egyezést $\rho = 1$ jelenti. Ha ρ kisebb, mint egy adott küszöbérték (ρ_0), akkor az adott oldal büntetésének mértékét a $(\rho_0 - \rho)$ különbség határozza meg.

Egy adott i oldal büntetésének kiszámításához azokat a j támogató oldalakat használtuk, amelyekre $\text{Support}_{i,j} > 0$. Azokat az oldalakat nem büntetjük, amelyeket támogató oldalak száma kisebb, mint $n_0 = 1000$.

A SpamRank algoritmus utolsó, 3. fázisában kapjuk az egyes oldalak végső SpamRank értékét. Ezt az egyes oldalak büntetéseiből álló vektoron kiszámolt perszonalizált PageRank vektorként kapjuk.

A SpamRank algoritmus az legelsőként publikált spamszűrési eredmények egyike. A konferencia változatot [C1] eddig 54-szer hivatkozták. Az eredmények Benczúr Andrással, Sarlós Tamással és Uher Mátéval közösek. Az eredményekhez az algoritmus részleteinek kidolgozásával és a kiértékeléssel járultam hozzá.

1.1. tézis. A SpamRank módszer [J2, C1]

A méréseink alapján a SpamRank módszer képes különbséget tenni spam és nem spam oldalak között.

A méréseket a `.de` doménhez tartozó 31 millió oldalból álló adathalmazon teszteltük, melyhez egy 1000 oldalból álló felcímkézett minta állt a rendelkezésünkre.

2. téziscsoport: Hivatkozás-alapú hasonlóság

Egy spam oldal általában nem izoláltan jelenik meg. A spammerek nagy számú, többé-kevésbé hasonló Weboldalakból álló farmokat generálnak, amelyeknek a célja, hogy félrevezessék a keresőrendszerek

rangsoroló algoritmusait. A mesterségesen létrehozott jelleg és spam oldalak közti erős kapcsolat lehetőséget teremt arra, hogy felismerjük őket. Az ajánló rendszerek és P2P hálózatok területeiről eredő ötlet a bizalom és bizalmatlanság továbbterjesztése (trust/distrust propagation), amely során már ismert spam és becsületes oldalakból kiindulva a hiperhivatkozások mentén más oldalakat is becsületes vagy spam oldalként azonosíthatunk. A legtöbb ilyen módszer PageRank-szerű továbbterjesztést használ. Mi ezekkel szemben a következő, hivatkozás-alapú hasonlóságot kiaknázó módszereket vizsgáltunk meg.

- A *Kocitáció* ($\mathrm{coc}(u, v)$) azon oldalak száma, amelyek mind u-ra, mind v-re is hivatkoznak.
- A *Companion* algoritmus [4] a HITS algoritmust [3] egy u Weboldal környezetében használja, és legmagasabb tekintély (authority) értékkel rendelkező oldalakat tekinti u-hoz hasonlónak.
- A Jeh és Widom [8] által javasolt a *SimRank* algolritmus elve az, hogy két oldal akkor hasonló, ha hasonló oldalak hivatkoznak rájuk. A SimRank a PageRankhez hasonlóan definiálható:

$$\mathrm{Sim}^{(0)}(u_1, u_2) = \begin{cases} 1 & \text{ha } u_1 = u_2 \\ 0 & \text{különben;} \end{cases} \tag{1}$$

$$\mathrm{Sim}^{(i)}(u_1, u_2) = \begin{cases} (1-c) \cdot \frac{\sum \mathrm{Sim}^{(i-1)}(v_1, v_2)}{d^-(u_1) \cdot d^-(u_2)} & \text{ha } u_1 \neq u_2, \\ 1 & \text{ha } u_1 = u_2. \end{cases} \tag{2}$$

- A Webgráfot *Szinguláris érték felbontás* segítségével egy alacsony dimenziójú térbe vetítettük, ahol az euklideszi távolságot használtuk a hasonolóság mértékeként.

Minden ismeretlen oldalhoz a következő spam-hasonlósági értékeket határoztuk meg:

- Spam Ratio (SR): A spam aránya a felcímkézett hasonló oldalak között ($s/(s+h)$).
- Spam over Non-spam (SON): A spam és a becsületes oldalak aránya az összes hasonló oldal között (s/h).
- Spam Value Ratio (SVR): A spam oldalak hasonlósági értékeinek összege osztva az összes felcímkézett oldalak hasonlósági értékeinek összegével ($s^*/(s^* + h^*)$).
- Spam Value over Non-spam Value (SVONV): A spam oldalak hasonlósági értékeinek összege osztva a becsületes oldalak hasonlósági értékeinek összegével ($s^*/h-^*$).

Az eredmények [C3] Benczúr Andrással és Sarlós Tamással közösek. Az eredményekhez az algoritmusok megtervezésével és hatékony implementálásával, valamint a kiértékeléssel járultam hozzá.

2.1. tézis. Hivatkozás alapú hasonlóság szerinti továbbterjesztés [C3]

Fő eredményként azt állítjuk, hogy magas felidézés mellett a link alapú hasonlóság szerinti továbbterjsztés hatékonyabb, mint a PageRank alapú. A hasonlósági mértékek közül a Cocitation a leghatékonyabb.

A módszert két adathalmazon is teszteltük: egy, a `.de` doménhez tartozó 31 millió oldalból álló adathalmazon, valamint a `.ch` domén hasonló méretű adathalmazán.

3. téziscsoport: Gráf alapú klasszifikátor-kombináció (Graph Stacking)

Kísérleteink alapján az előző téziscsoportbeli hasonlóságmértékek alkalmazhatók félig felügyelt (semi-supervised) tanulási módszereknél is. Ezek a tanulási módszerek is arra a tapasztalatra építenek, hogy egy Weboldal hasonló a szomszédaihoz. Méréseink alapján a hivatkozás-hasonlósággal súlyozott gráf alapú klasszifikátor-kombináció *stacked graphical classification* növeli a spam klasszifikácó pontosságát. A legjobb eredményt a kocitációval sikerült elérni, valamint azt is megállpítottuk, hogy a kettőnél több lépésés szomszédségot nem érdemes vizsgálni.

A spam klasszifikáció mellett egy egészen más típusú adathalmazon is kiértékeltük az előbbi módszert: egy telefonhívások által adott hálózaton, amelyet a telefon előfizetők bizonyos adatai egészítettek ki. A tanulási feladat célja megjósolni, hogy melyik ügyfél fogja az előfizetését lemondani és egy másik szolgáltatóhoz átpártolni (churn). Ha első látásra nem is, de ez a probléma hasonlít a Web spam kereséshez: a Webgráf helyett telefonhívás-gráfot, a Weboldalak tartalma helyett előfizetői, esetleg demográfiai adatokat használunk. A spamnél megfigyelt jelenség itt is megtalálható: azok az előfizetők nagyobb eséllyel váltanak szolgáltatót, akiknek ismerősei már váltottak.

Az algoritmus során a klasszifikációs algoritmust több lépésben megismételjük úgy, hogy az előző klasszifikáció eredménye és a hasonlógi mértékek alapján minden csúcsra kiszámolunk egy spam-hasonlósági mértéket, amelyet a következő iterációban új jellemzőként használunk. Minden ismeretlen u csúcsra kiválasztjuk azt a k darab csúcsot, amelyek a legnagyobb hasonlósággal rendelkeznek, és ezekre az előző pontnál leírt spam mértékeket (Spam Ratio, Spam over Non-spam, Spam Value Ratio, Spam Value over Non-spam Value) számítjuk ki. Az előző téziscsoport módszereihez képest módosítás, hogy most a kezdetben ismeretlen csúcsokat is figyelmbe vesszük az előző iteráció által hozzárendelt értéknek megfelelően. A kiválasztott csúcsok minden v eleme vagy $p(v)$ jósolt valószínűséggel spam, vagy pontosan tudjuk, hogy spam vagy becsületes. Az utóbbi esetben $p(v)$-t 0-nak ill. 1-nek definiáljuk. A spam mértékek definíciójában szereplő s és h értékeket a spam illetve becsületes csúcsok $p(v)$ illetve $1 - p(v)$ értékeinek öszegeként definiálhatjuk. A leghasonlóbb oldalakat az s^* és h^* értékek $w(uv)$-vel súlyozott összegeként kapjuk.

Az eredmények [C4] Benczúr Andrással, Siklósi Dáviddal és Lukács Lászlóval közösek. Az hozzájárulásom a gráfos kompináción alapuló jellemzők definiálása, implementálása, a hasonlósági mértékek

megtervezése és implementálása.

3.1. tézis. Gráf alapú klasszifikátor-kombináció (Graph Stacking) Web spam szűrésben [C4]

Méréseink alapján a hivatkozás-hasonlósággal súlyozott gráf alapú klasszifikátor-kombináció (Graph Stacking) növeli a spam klasszifikáció pontosságát.

A módszert a WEBSPAM-UK-2006 [2] adathalmazon teszteltük.

3.2. tézis. Gráf alapú klasszifikátor-kombináció (Graph Stacking) lemorzsolódás (churn) előrejelzésére [C4]

Méréseink alapján a hivatkozás-hasonlósággal súlyozott gráf alapú klasszifikátor-kombináció (Graph Stacking) növeli a churn klasszifikáció pontosságát.

A módszert egy magyar telefonszolgáltató adatain teszteltük.

4. téziscsoport: Nyelvmodell különbözőség

Mishne [10] blogok hozzászólásait elemezve megmutatta, hogy a szavak eloszlása (unigram nyelvmodell) jól használható a spam jellegű hozzászólások kiszűrésére. Megmutattuk, hogy hasonló módszer nagyobb méretekben a Web spam szűrésre is alkalmazható.

A [10] cikk módszeréhez hasonlóan egy D Weboldal nyelvmodelljét a következőképp definiáljuk:

$$p(w|D) = \lambda \frac{tf(w,D)}{\sum_{v \in D} tf(v,D)} + (1-\lambda) \frac{tf(w,C)}{\sum_{v \in C} tf(v,C)}, \tag{3}$$

ahol C az összes Weboldalból álló korpuszt, w pedig a D egy szavát jelöli.

Módszerünk első lépése az, hogy megmérjük minden egyes hiperhivatkozás esetén az azt tartalmazó illetve a hiperhivatkozás által mutatott két Weboldal közti nyelvi különbséget. A távolságot a [10] cikkhez hasonlóan a Kullback-Leibler divergencia alapján határozzuk meg:

$$KL(A\,||D) = \sum_{w} p\,(w|A) \log \frac{p\,(w|A)}{p\,(w|D)}. \tag{4}$$

Sajnos a KL érték kiszámolása minden hiperhivatkozással összekötött Weboldalpárra túl sok időt venne igénybe, ezért két hasonló jellegű, de egyszerűbben számolható mennyiséget vizsgálunk meg. A hivatkozást tartalmazó Weboldal teljes tartalma helyett csak a megfelelő anchor szövegét használjuk, és ezt hasonlítjuk össze a tartalmazó illetve az általa mutatott Weboldal tartalmával. A módszer hatékonyságát javítja, ha az anchor szövegét még kiegészítjük környező szavakkal.

[10] szerint a Kullback-Leibler divergencia értékek normál eloszlások keverékét alkotják. Ha minden hivatkozás egyformán viselkedne, akkor a KL értékek nagyjából normál eloszlást mutatnának, mert a KL definíciójában szereplő valószínűségi változók különböző szavakhoz tartoznak, és a szavak elég függetlenül viselkednek ahhoz, hogy az eredmény normál eloszlású legyen. Azonban ha vannak olyan spam jellegű hiperhivatkozások is, amelyeknek az anchor szövege jelentősen eltér az általa mutatott, vagy az őt tartalmazó Weboldal tartalmáttól, akkor ezeknek a KL értékei egy becsületes hiperhivatkozások átlagértékénél magasabb érték körül fognak csoportosulni.

Az algoritmusunkban egy küszöbérték feletti KL értékkel rendelkező hiperlinket gyanús tekintjük, és az egyes oldalak spam mértékét a gyanús hiperlinkeken számolt PageRank értékeként definiáljuk.

Az eredmények [P1] Benczúr Andrással és Bíró Istvánnal közösek. Az én hozzájárulásom az algoritmus alkalmazhatóságának kiterjesztése nagyméretű adatokra és a kiértékelés.

4.1. tézis. Nyelvmodell különbözőség alkalmazása spam szűrésre [P1]

Megmutattuk, hogy a nyelvmodell különbözőség használható Web spam szűrésre, valamint egyéb, rossz minőségű hiperhivatkozások kiszűrésére.

A méréseket a `.de` doménhez tartozó 31 millió oldalból álló adathalmazon teszteltük, melyhez egy 1000 oldalból álló felcímkézett minta állt rendelkezésünkre.

5. téziscsoport: Kereskedelmi szándék

A spam elsődleges célja anyagi haszonszerzés [6]. A korábbi tartalom alapú spam szűrési modszerek főleg a spam oldalak ismétlődő mintázatait használták ki, ezzel szemben a mi módszerünk megpróbálja tartalom szemantikáját is figyelembe venni, elsősorban a kereskedelmi jelleg felismerésére koncentrálva.

A következő értékeket, tulajdonságokat vizsgáltuk meg:

- A Weboldalak Online Commercial Intention (OCI) értéke (Microsoft adCenter Labs Demonstration.)
- A Yahoo! Mindset kereskedelmi/nem kereskedelmi klasszifikációja.
- Google AdWords által ajánlott hirdetési keresőszavak és azok értéke.
- Google AdSense hirdetések eloszlása egy Website-on.
- A spam sikeresség, azaz a találatok listájában elfoglalt hely bizonyos keresések esetén (a saját fejlesztésű keresőmotorunkon mérve).

Ezeket a jellemzőket kombináltuk egy akkoriban nyilvánosan elérhetővé tett tartalom és hivatkozás alapú leíró adathalmazhoz, és az így kapott tanító adaton C4.5 döntési fa alapú klasszifikációt alkalmaztunk.

Az eredmények [C5] Benczúr Andrással, Bíró Istvánnal és Sarlós Tamással közösek. Az eredményekhez a kiértékeléssel, a jellemzők kombinálásával, a spam sikeresség, és a Google AdWords jellemzők definíciójával és implementációjával járultam hozzá.

5.1. tézis. A kereskedelmi szándék felismerésének alkalmazása spam szűrésrei [C5]

A kereskedelmi szándékot kifejező tulajdonságok javítják a spam klasszifikáció minőségét.

A módszert a WEBSPAM-UK2006 adathalmazon teszteltük, ahol 3%-kal meg tudtuk javítani a csak a többi nyilvánosan elérhető adatokat használó klasszifikátor teljesítményét. Módszerünk F mértékre nézve első, AUC mértékre pedig holtversenyben második helyezést ért el.

3. Egyéb eredmények

A spam szűrési módszereken kívüli egyéb, főleg a Webes kereséssel kapcsolatos eredményeim:

- Magyar nyelvű keresőmotor fejlesztése [J6, P2, H1, H2, C9].
- Képi információk kombinálása szövegalapú kereséssel [J5, C8].
- Spektrálklaszterezés nagy méretű gráfokon és alkalmazásai ajánlórendszerekben és telekommunikációs hálózatokban [J4, C6, C7, H3].

Hivatkozások

[1] A.-L. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: the topology of the word-wide web. *Physica A*, 281:69–77, 2000.

[2] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2):11–24, December 2006.

[3] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the Web's link structure. *Computer*, 32(8):60–67, 1999.

[4] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. In *Proceedings of the 8th World Wide Web Conference (WWW)*, pages 1467–1479, 1999.

[5] S. Dill, R. Kumar, K. S. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Transactions on Internet Technology*, 2(3):205–223, 2002.

[6] Z. Gyöngyi and H. Garcia-Molina. Spam: It's not just for inboxes anymore. *IEEE Computer Magazine*, 38(10):28–34, October 2005.

[7] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2):11–22, 2002.

[8] G. Jeh and J. Widom. SimRank: A measure of structural-context similarity. In *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 538–543, 2002.

[9] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proceedings of the 41st IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 1–10, 2000.

[10] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Chiba, Japan, 2005.

[11] G. Pandurangan, P. Raghavan, and E. Upfal. Using PageRank to Characterize Web Structure. In *Proceedings of the 8th Annual International Computing and Combinatorics Conference (COCOON)*, pages 330–339, Singapore, 2002.

[12] A. Singhal. Challenges in running a commercial search engine. In *IBM Search and Collaboration Seminar 2004*. IBM Haifa Labs, 2004.

Publikációk

Külföldi idegen nyelvű folyóiratcikkek

[J1] Dániel Fogaras, Balázs Rácz, Károly Csalogány, and Tamás Sarlós. Towards Scaling Fully Personalized PageRank: Algorithms, Lower Bounds, and Experiments. *Internet Mathematics* 2(3):333-358, 2005.

[J2] András A. Benczúr, Károly Csalogány, Tamás Sarlós, and Máte Uher. SpamRank – Fully Automatic Link Spam Detection. To appear in *Information Retrieval*.

[J3] M. Kurucz, A. Benczúr, K. Csalogány, and L. Lukács. Spectral Clustering in Social Networks. *Lecture Notes In Artificial Intelligence*, pages 1–20, 2009.

[J4] M. Kurucz, D. Siklósi, L. Lukács, A. A. Benczúr, K. Csalogány, and A. Lukács. Telephone call network data mining: A survey with experiments. In Bolyai Society Mathematical Studies, Vol. 18., B. Bollobás, R. Kozma, D. Miklós, eds., *Handbook of Large-Scale Random Networks, published by Springer Verlag in conjunction with the Bolyai Mathematical Society of Budapest*, 2008.

[J5] A. Benczúr, I. Bíró, M. Brendel, K. Csalogány, B. Daróczy, and D. Siklósi. Multimodal Retrieval by Text–Segment Biclustering. In *ADVANCES IN MULTILINGUAL AND MULTIMODAL INFORMATION RETRIEVAL*. 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Revised Selected Papers. Lecture Notes in Computer Science Vol. 5152 (2008).

[J6] P. Schönhofen, A. A. Benczúr, I. Bíró, and K. Csalogány. Cross-language retrieval with wikipedia. In *ADVANCES IN MULTILINGUAL AND MULTIMODAL INFORMATION RETRIEVAL*. 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Revised Selected Papers. Lecture Notes in Computer Science Vol. 5152 (2008).

Nemzetközi konferencia-kiadványban megjelent cikkek

[C1] András A. Benczúr, Károly Csalogány, Tamás Sarlós, and Máte Uher. SpamRank – Fully Automatic Link Spam Detection. In *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), held in conjunction with WWW2005*, 2005.

[C2] Tamás Sarlós, András A. Benczúr, Károly Csalogány, Dániel Fogaras, and Balázs Rácz. To Randomize or Not To Randomize: Space Optimal Summaries for Hyperlink Analysis In *Proceedings of the 15th International World Wide Web Conference (WWW)*, pages 297-306, 2006.
Full version available at http://datamining.sztaki.hu/www/index.pl/publications-en.

[C3] A. A. Benczúr, K. Csalogány, and T. Sarlós. Link-based similarity search to fight Web spam. In *Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), held in conjunction with SIGIR2006*, 2006.

[C4] K. Csalogány, A. Benczúr, D. Siklósi, and L. Lukács. Semi-Supervised Learning: A Comparative Study for Web Spam and Telephone User Churn. In *Graph Labeling Workshop in conjunction with ECML/PKDD 2007*, 2007.

[C5] A. A. Benczúr, I. Bíró, K. Csalogány, and T. Sarlós. Web spam detection via commercial intent analysis. In *Proceedings of the 3th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), held in conjunction with WWW2007*, 2007.

[C6] M. Kurucz, A. A. Benczúr, and K. Csalogány. Methods for large scale svd with missing values. In *KDD Cup and Workshop in conjunction with KDD 2007*, 2007.

[C7] M. Kurucz, A. A. Benczúr, K. Csalogány, and L. Lukács. Spectral clustering in telephone call graphs. In *WebKDD/SNAKDD Workshop 2007 in conjunction with KDD 2007*, 2007.

[C8] A. Benczúr, I. Bíró, M. Brendel, C. Károly, B. Daróczy, and D. Siklósi. Cross-modal retrieval by text and image feature biclustering. In *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 2007.

[C9] P. Schönhofen, A. A. Benczúr, I. Bíró, and K. Csalogány. Performing cross-language retrieval with Wikipedia. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, Sept. 2007.

Nemzetközi konferencia-kiadványban megjelent poszterek

[P1] A. A. Benczúr, I. Bíró, and K. Csalogány. Detecting nepotistic links by language model disagreement. In *Proceedings of the 15th International World Wide Web Conference (WWW)*, 2006.

[P2] András A. Benczúr, Károly Csalogány, Dániel Fogaras, Eszter Friedman, Tamás Sarlós, Máte Uher, and Eszter Windhager. Searching a Small National Domain – a Preliminary Report. In *Poster Proceedings of the 12th International World Wide Web Conference (WWW)*, 2003.

[P3] András A. Benczúr, Károly Csalogány, and Tamás Sarlós. On the Feasibility of Low-rank Approximation for Personalized PageRank. In *Poster Proceedings of the 14th International World Wide Web Conference (WWW)*, pages 972–973, 2005.

Magyar nyelvű publikációk

[H1] András A. Benczúr, Károly Csalogány, Dániel Fogaras, Eszter Friedman, Balázs Rácz, Tamás Sarlós, Máte Uher, and Eszter Windhager. Magyar nyelvű tartalom a világhálón (Hungarian Content on the WWW). *Információs Társadalom és Trendkutató Központ Kutatási Jelentés* 26:48-55, 2004.

[H2] András A. Benczúr, István Bíró, Károly Csalogány, Balázs Rácz, Tamás Sarlós, and Máte Uher. PageRank és azon túl: Hiperhivatkozások szerepe a keresésben (PageRank and Beyond: The Role of Hyperlinks in Search). *Magyar Tudomány*, 2006.

[H3] Miklós Kurucz, László Lukács, Dávid Siklósi, András A. Benczúr, Károly Csalogány, András Lukács Kapcsolatok és távolságok: a hazai vezetékes hívás-szokások elemzése. (Contacts and Distances: Analysis of Hungarian Landline Telephone Calls). *Magyar Tudomány*, 2009/6.